Rule-based Automatic Text Simplification for German

Suter, Julia ; Ebling, Sarah ; Volk, Martin

Rule-based Automatic Text Simplification for German

Julia Suter **Sarah Ebling** **Martin Volk**
Institute of Computational Linguistics, University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
suter@cl.uni-heidelberg.de, {ebling|volk}@cl.uzh.ch

Abstract

Automatic text simplification is capable of rendering texts comprehensible and accessible to persons with difficulties in reading and processing written language. In this paper, we report on the development of a rule-based automatic text simplification system for German. We show that the complexity of the output of our system is comparable to that of simplifications produced by a human.

1 Introduction

Simplified language aims to make texts comprehensible and accessible to persons with difficulties in reading and processing written language.[1] It is aimed not just at cognitively impaired persons but also at functionally illiterate and deaf persons, persons suffering from dementia and other neurodegenerative diseases, and immigrants.

Simplified language is characterized by reduced lexical and syntactic complexity, the addition of explanations for difficult words, and a clearly structured layout. Text in simplified language is usually obtained by simplifying a text written in standard language. By definition, simplification should not alter the meaning and informative value of a standard-language text (Coster and Kauchak, 2011a); this is what distinguishes it from other text-to-text generation tasks such as text compression.

Automatic text simplification, the process of automatically producing a simplified text, has only recently become an established research topic. It offers the potential of both increasing readability and comprehensibility for humans and improving processability for machines. As an example of the latter, text simplification as a preprocessing step can increase performance of natural language processing tasks such as parsing, machine translation, information retrieval, and text summarization (Chandrasekar et al., 1996).

Automatic text simplification systems have been developed for languages such as English, Swedish, and Portuguese. While tools exist for *detecting* complex structures in German texts,[2] to the best of our knowledge, no system exists for automatically *simplifying* these structures. On a more general level, Matausch and Nietzio (2012) state that "plain language is still underrepresented in the German speaking area and needs further development".

In this paper, we report on the development of a rule-based automatic text simplification system for German. Our approach builds on simplification rules extracted from guidelines for simplified German. We show that the complexity of the output of our system is comparable to that of simplifications produced by a human.

The remainder of this paper is structured as follows: Section 2 discusses the guidelines we used as a basis for our German simplification system (Section 2.1) as well as previous approaches to automatic text simplification for languages other than German (Section 2.2). Section 3 introduces our simplification system, discussing the resources used (Section 3.1) and simplification method applied (Section 3.2) as well as presenting an evaluation (Section 3.3) and discussing the results thereof (Section 3.4).

2 Simplified German

2.1 Guidelines

Guidelines specifying the character set, vocabulary, linguistic structures, and layout permitted for

[1] Related terms are *plain language*, *simple language*, or *easy-to-read language*. The term *simplified language* is used throughout this paper to emphasize the fact that the underlying concept is by no means standardized, as will become obvious in Section 2.1.

[2] An example is the LanguageTool (https://www.languagetool.org/de/leichte-sprache/).

simplified language are essential for systematically simplifying a standard-language text. For simplified German, four well-known guidelines exist: the guidelines by Inclusion Europe (2009),Netzwerk Leichte Sprache (2009), the BITV 2.0 rule set (Bundesministerium der Justiz und für Verbraucherschutz, 2011), and the guidelines by Maaß (2015).

Simplified language is still a young phenomenon, with profound research on the concept, its target groups, and guidelines still ongoing. No standardized version of simplified German exists. Accordingly, the four guidelines introduced above do not always agree on the best way to simplify complex language. The guidelines by Maaß (2015) provide the most coherent, linguistically motivated recommendations for transforming standard German into simplified German. We therefore based our work on these guidelines, well aware that some of our simplification rules might need adjustment at a later stage as more research is carried out in the respective area.

The guidelines by Maaß (2015) are divided into five categories according to the level of language they concern: character level, word level, sentence level, text level, and layout.

2.1.1 Character level

The character set of simplified German contains the letters of the German alphabet, an extension of the Latin alphabet with umlaut vowels (ä, ö, ü). In addition, digits and the special characters . ? ! „ “ : · are permitted. Other special characters such as the paragraph symbol (§) or dollar sign ($) are not allowed. The comma is not part of the inventory of simplified German according to Maaß (2015), as subordinate clauses and enumerations, which are typically introduced by or contain commas, should not be used. Numbers should be written as digits rather than words, with the exception of the indefinite article *ein* (‘a’), which is to be written as a word to prevent ambiguity with the cardinal number 1. Since compounds are productive in German, Maaß (2015) proposes the use of a typographical device called *Mediopunkt* (‘center dot’) to visually segment compounds, e.g., *Unfall·versicherung* (‘accident insurance’). Other guidelines suggest using hyphens in compounds; however, this requires capitalization of the compound segments and can lead to non-standard spelling.

2.1.2 Word level

Simplified language may contain only simple, short, and well-known words. Technical terms, foreign words, and abbreviations should be avoided, though common acronyms such as *CD* may be used. In cases where a difficult word is unavoidable, the word should be explained in simple terms. So far, no vocabulary list for simplified German exists.

2.1.3 Sentence level

Each sentence in simplified language should only contain one piece of information. Therefore, coordinate and subordinate clauses should be transformed into independent main clauses. Main clauses should preferably contain subject-verb-object (SVO) word order, active voice, and present or past perfect tense. Negations, nominal style, and metaphors should be avoided. Rare morphological forms may be unknown to inexperienced readers, so genitive case, subjunctive mood, and past simple tenses need to be eliminated.

2.1.4 Text level

In simplified language, consistency is given preference over style: word repetition and linear syntactic structures are encouraged, even though this conflicts with stylistic conventions common in standard language. Synonyms and third-person pronouns should be replaced with their antecedent noun phrases. Indirect speech is to be rephrased as direct speech. Additionally, a text may be enhanced with examples and explanations. Pictures, charts, and graphics should only be used if they are meaningful and appropriate for the target readership.

2.1.5 Typography and layout

Simplified German is always displayed one sentence per line. If a sentence takes up more than one line, it should be segmented at syntactic phrase boundaries. Text should be set in a large sans-serif font type and structure emphasized by means of headlines and indentations.

2.2 Automatic text simplification

Automatic text simplification can be performed using rule-based or corpus-based (mostly statistical) approaches. Rule-based automatic text simplification systems have been developed, e.g., for English, Swedish, French, Spanish, and Portuguese. These systems perform, among other tasks, lexical simplification (Kandula et al., 2010; Paetzold and

Specia, 2015), explanation generation (Watanabe et al., 2010), and syntactic simplification such as splitting long coordinate and subordinate phrases, rephrasing appositives and relative clauses (Aluísio and Gasperin, 2010), resolving third-person pronouns (Siddharthan, 2006), changing passive to active voice, and rearranging irregular word order (Rennes and Jönsson, 2015).

Corpus-based approaches have taken, e.g., the form of simplification via statistical machine translation (SMT) in the past (Coster and Kauchak, 2011a; Coster and Kauchak, 2011b; Specia, 2010; Stymne et al., 2013). Klaper et al. (2013) created a parallel German/Simple German corpus containing 70,000 tokens for use in SMT. However, they did not train an SMT system.

3 Rule-based simplification for German

We decided to follow a rule-based approach to text simplification for a number of reasons. Most importantly, statistical approaches require large amounts of data, something that is not available for simplified German to date. The parallel corpus by Klaper et al. (2013) mentioned in Section 2.2 cannot be expected to be sufficiently large to train an SMT system that works reasonably well. Secondly, if text simplification is used as an assistive technology, it is essential that it produces accurate results. Meaningless paraphrasings produced by an SMT system or other statistical methods can be even more confusing than the original (non-simplified) text (Shardlow, 2014). Finally, the guidelines by Maaß (2015) suggest simplification steps that can only be achieved through syntactic transformation rules. Statistical approaches are not well equipped to handle simplifications that require syntactic reordering, morphological transformations, and insertions due to lack of explicit linguistic knowledge (Siddharthan, 2014).

3.1 Resources

Our system makes use of a number of external resources for German. For example, it is based on the output of syntactic parsing of the source text. We employ the hybrid dependency parser ParZu (Sennrich et al., 2009), which performs sentence segmentation and tokenization. For compound segmentation, we use the tool Gertwol, which returns all possible segmentations of a word and provides further morphological analysis (Haapalainen and Majorin, 1995). For selecting the best segmentation, we implemented the algorithm suggested by Volk (1999), which ranks compound candidates according to their internal complexity of composition and derivation boundaries.

To retrieve abbreviations and their corresponding full forms, we extracted a list of 405 abbreviations and 278 acronyms from Wikipedia.[3] For verb conjugation, we rely on a web service that provides conjugation tables for most German verbs in all tenses and modes.[4] Nominals are inflected using CanooNet, an online dictionary that contains more than 250,000 manually checked German word entries.[5] Short definitions of difficult words are extracted from Hurraki, a Wiki-style encyclopaedia for simplified German consisting of more than 2,400 articles.[6]

3.2 Method

We implemented a subset of the rules described in Section 2.1 to perform automated simplification on the character, word, sentence, and text level and to adjust the layout of the output. The architecture of our system is shown in Figure 1. Following preprocessing and syntactic parsing of the source text (cf. Section 3.1), the simplification rules are applied sentence by sentence.

3.2.1 Character- and word-level rules

Prior to the parsing step, parentheses and their enclosed contents are removed from the source text, and abbreviations are expanded to their full forms. Both steps simplify the text and improve parsing performance. Following the parsing step, numbers written as words and special characters are replaced by digits and appropriate word substitutions using manually created dictionaries. All nouns longer than five characters that are not proper names are examined as to whether they are compounds; if this is the case, they are split using the *Mediopunkt* (cf. Section 2.1.1). Sample character- and word-level transformations are shown in Example 1.

(1) **German**
Prof. Müller kauft sich den siebten Band (den letzten) seiner Lieblingskrimireihe für 8$ 50¢ inkl. MwSt.

[3] https://de.wikipedia.org/wiki/Portal:Abkürzungen/Gebräuchliche_Abkürzungen
[4] http://www.verbformen.de/
[5] http://www.canoo.net/
[6] http://hurraki.de/

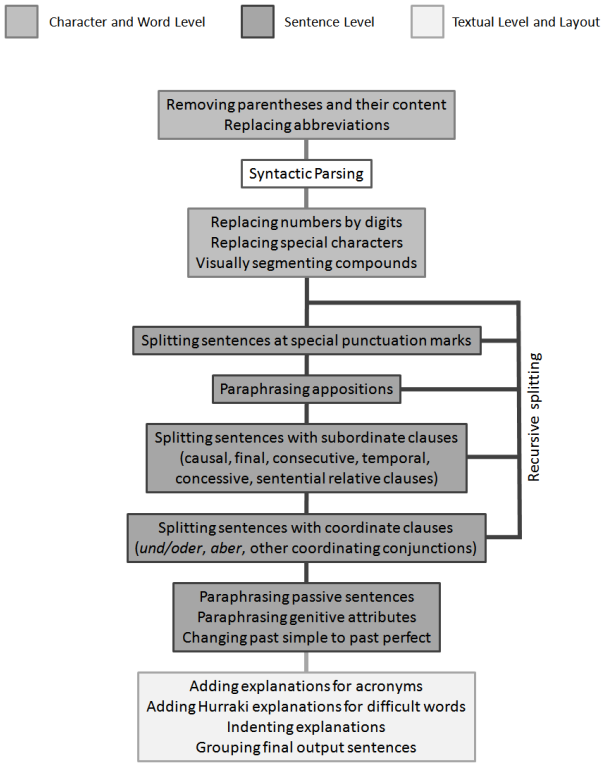

Figure 1: Architecture of rule-based text simplification system.

'Prof. Müller buys the seventh volume (the last one) of his favorite crime novel series for 8$ 50¢, incl. VAT.'

Simplified German
Professor Müller kauft sich den 7. Band von seiner Lieblings·krimi·reihe für 8 Dollar 50 Cent inklusive Mehrwert·steuer.

'Professor Müller buys the 7th volume of his favorite crime novel series for 8 dollars 50 cents, including value-added tax.'[7]

3.2.2 Sentence-level rules

On the sentence level, a series of syntactic simplification rules are executed. These rules split and/or rephrase the sentences. Syntactic simplification is applied recursively. The individual simplification rules are either executed once per iteration or are triggered by "keywords" (words or special characters), as described in what follows.

Syntactic simplification begins by looking for semicolons and dashes and splitting sentences at these characters. Sentences are also split after colons if the segment after the colon is a complete sentence and not just an enumeration.

Appositions are replaced by sentences in which the noun phrase referred to by the apposition forms the subject (*X*) and the apposition itself becomes the predicative noun (*Y*), yielding an *X is Y* structure (cf. Example 2).

(2) **German**
Der Artikel wurde von Dr. Meier, dem Leiter der Universitätsklinik, verfasst.
'The article was written by Dr Meier, head of the university hospital.'

Simplified German
Doktor Meier hat den Artikel verfasst.
Meier ist der Leiter von der Universitäts·klinik.
'Doctor Meier has written the article.
Meier is head of the university hospital.'

Rules for rephrasing subordinate clauses all have a similar structure: If a subordinate conjunction is found, the sentence is split at the conjunction and both resulting segments are edited and rephrased to form independent sentences. Suitable connectives that express the rhetorical relation are added to preserve the original meaning, and the correct word order is restored. For instance, sentences containing causal clauses, e.g., with *weil* ('because') or *da* ('since'), are split into two main clauses, and the latter clause is complemented with the connective *deshalb* or *denn* ('thus') to maintain the causal relation (cf. Example 3).

(3) **German**
Weil der Gastgeber noch nicht da ist, müssen die Gäste warten.
'Since the host is not there yet, the guests have to wait.'

Simplified German
Der Gastgeber ist noch nicht da.
Deshalb müssen die Gäste warten.
'The host is not there yet.
Therefore, the guests have to wait.'

Concessive clauses with subjunctions like *obwohl* ('although') are rephrased using the connective *trotzdem* ('however') (cf. Example 4). Consecutive clauses starting with *sodass* ('so that')

[7] Note that different from the English acronym ‚VAT', the German abbreviation ‚MwSt.' is always expanded to its full form ‚Mehrwertsteuer' in reading or speaking.

are rephrased using *deshalb* ('therefore'), possibly shifting the meaning slightly but essentially retaining the information. We found acceptable connectives for rephrasing temporal clauses in *nachdem* ('after'), *bevor* ('before'), *seit* ('since'), and *während* ('while'), yet we could not find a suitable solution for the conjunction *als* ('when/as'). Final clauses are rephrased using the modal verb *wollen* ('want') and the connective *deshalb* ('therefore'). Since the subject is not mentioned overtly in German final clauses containing *um zu* ('in order to'), it has to be retrieved from the main clause (cf. Example 5).

There is no general way of simplifying relative clauses, so we focused on sentential relative clauses, which do not refer to the preceding noun but to the whole sentence or clause. Such sentences can be split at the pronominal adverb, which is then replaced by its cataphoric adverb (cf. Example 4).

(4) **German**
Obwohl er seine Rechnungen immer pünktlich bezahlte, bekam er eine Mahnung, worüber er sich sehr ärgerte.
'Although he always paid his bills on time, he received a reminder, which really bothered him.'

Simplified German
Er hat seine Rechnungen immer pünktlich bezahlt.
Trotzdem hat er eine Mahnung bekommen.
Darüber hat er sich sehr geärgert.
'He has always paid his bills on time.
However, he has received a reminder.
This has really bothered him.[8]

(5) **German**
Um den Text verständlicher zu machen, verwenden wir nur einfache Wörter.
'To make the text easier to understand, we only use simple words.'

Simplified German
Wir wollen den Text verständlicher machen.
Deshalb verwenden wir nur einfache Wörter.
'We want to make the text easier to understand.
Therefore, we only use simple words.'

Coordinate clauses are split at coordinating conjunctions (e.g. *und* ('and'), *oder* ('or'), *aber* ('but'), *dennoch* ('however')). If the second resulting clause is elliptic, the missing subject or predicate is retrieved from the previous clause and the subject shortened, i.e., adjectives, genitive attributes, and prepositional phrases are removed. We allowed for sentences to start with *und* ('and') and *oder* ('or') to emphasize that they are linked to the previous sentence.

(6) **German**
Der junge Beamte an der Grenze überprüft die Reisepässe und kontrolliert das Gepäck der Fluggäste.
'The young officer at the border checks the passports and examines the passengers' luggage.'

Simplified German
Der junge Beamte an der Grenze überprüft die Reise·pässe.
Und der Beamte kontrolliert das Gepäck von den Flug·gästen.
'The young officer at the border checks the passports.
And the officer examines the luggage of the passengers.'

If a passive construction is detected, our system retrieves the grammatical agent indicated by a prepositional phrase starting with *von* ('by'), the object (the subject of the passive phrase), and the action verb (past participle) and generates a sentence in active voice. If the agent is not mentioned, we use the impersonal pronoun *man* ('one') as subject in the active-voice sentence (cf. Example 7). Although impersonal language should be avoided, we decided to accept the pronoun *man* when resolving passive constructions without explicit agent, as it is likely to be less difficult than the original passive construction. To rephrase genitive attributes, the entire attribute is transformed into dative case and complemented with the preposition *von* ('of') (cf. Example 6).

(7) **German**
Der Dieb wurde von der Polizei gefasst. Er wurde in Handschellen abgeführt.
'The thief was arrested by the police.
He was taken away in handcuffs.'

Simplified German
Die Polizei hat den Dieb gefasst.
Man hat ihn in Hand·schellen abgeführt.
'The police has arrested the thief.
One/they have taken him away in handcuffs.'

[8] Note that present perfect tense fulfills a slightly different function in German than it does in English.

If the sentence is in simple past, the tense is changed to past perfect. The auxiliary verb *sein* ('be') or *haben* ('have') is conjugated accordingly; the past participle is simply added at the end of the sentence (cf. Example 7). This works well, since the sentence is already highly simplified and shortened at this point. Auxiliary and modal verbs remain in past simple tense because they are well-known forms and their past perfect use is often deemed unnatural (Maaß, 2015).

3.2.3 Text-level and layout rules

Simplified language requires explanations for difficult words. We regard as difficult vocabulary acronyms (derived from Wikipedia) and words that are explained in the Hurraki online dictionary (cf. Section 3.1). Acronyms are explained after their first occurrence in the text but are not expanded like abbreviations, to avoid long and difficult words. For non-trivial words with a Hurraki entry, the short Hurraki definition is retrieved and inserted into the text. Some Hurraki explanations do not conform to the guidelines of Maaß (2015); we refrained from modifying them. To mark added explanations automatically and make the text more readable, explanations are indented (cf. Example 8). When printing the final simplified text, all sentences resulting from one original sentence are grouped together in a paragraph to emphasize which information belongs together.

(8) **German**
Andreas Meyer ist der Chef der SBB.
'Andreas Meyer is the director of SBB.'

Simplified German
Andreas Meyer ist der Chef der SBB.[9]
'Andreas Meyer is the director of SBB.'

SBB ist die Abkürzung für
Schweizerische Bundesbahnen.
Chef ist ein schwieriges Wort.
Hurraki erklärt es so:
Ein Chef ist im Betrieb der Vorgesetzte oder Verantwortliche.

'SBB is the abbreviation for Swiss Federal Railways.
Director is a difficult word.
Hurraki explains it as follows:
A director is the supervisor or responsible person in a company.'

[9] Since the parser does not recognize *der SBB* as a genitive attribute, it is not modified.

3.3 Evaluation

A common way of evaluating simplified texts is to apply readability metrics. Readability metrics typically assess one or multiple surface features such as word or sentence length. Well-known examples are the Flesch Reading Ease Score (Flesch, 1948) and the *Läsbarhetsindex* ('readability index', LIX) (Björnsson, 1968). Flesch Reading Ease measures word length in syllables and sentence length in words and delivers a score on a scale from 0 to 100, with higher scores indicating better readability. Flesch is frequently used to assess writings of students in U.S. grade schools. LIX computes the sum of the average sentence length and the ratio of long words (i.e., words with more than six letters). Like Flesch, the resulting score ranges between 0 and 100; however, with LIX, higher scores correlate with lower readability.

Metrics like Flesch and LIX are generally understood to cover only a part of what constitutes the readability of a text (Chall, 1958). Heimann Mühlenbock (2013) developed the more sophisticated SVIT model for assessing the readability of Swedish texts. Since the model is partly language-specific, we could not rely on it for evaluating the simplifications produced by our system. Addressing "the current problem in the text simplification community that there are no common standards and evaluation methodologies which would enable fair comparison of different ATS [automatic text simplification; the authors] systems" was the aim of a workshop at the Language Resources and Evaluation Conference.[10]

We evaluated the output of our system both quantitatively, by computing its LIX score (well aware of the shortcomings of this score), and qualitatively, by comparing it to a simplification produced by a person who had undergone the six-month *Leicht Lesen* ('easy-read') training offered for German by the *capito* network.[11]

3.3.1 Data

The evaluation text is a short article on the arrival of the Swiss team at the Special Olympics in Korea. It consists of 135 words in six sentences and features many aspects of standard language: long, difficult, and foreign words, exclamation marks, dashes and colons, appositions, long and elliptic sentences, coordinate clauses, one participle construction, final

[10] http://qats2016.github.io/index.html
[11] http://www.capito.eu/de/Leicht_Lesen/

LIX score	Description	Text type
<25	very easy	children's literature
25-30	easy	young adults' literature
30-40	standard	fiction and daily news
40-50	fairly difficult	informative texts, non-fiction
50-60	difficult	specialist texts
>60	very difficult	scientific texts

Table 1: Description of LIX scale (table from (Heimann Mühlenbock, 2013, p. 32).

clause, and sentential relative clause, passive constructions, genitives, and past simple forms. The text was chosen on the basis that it contains many complex structures. It had not been used to develop the system.

3.3.2 Quantitative evaluation

The human-simplified text contains 152 words in 16 sentences, the text resulting from our simplification consists of 146 words in 14 sentences (with added Hurraki explanations: 217 words in 24 sentences). The original (standard-language) version of the evaluation text has a LIX score of 53, which corresponds to the level of difficulty of specialist texts according to the classification shown in Table 1 (Heimann Mühlenbock, 2013). The LIX score of the human simplification is 35,[12] assigning a "standard level" of difficulty to the text, similar to fictional and daily news texts. The simplification generated by our system has a LIX score of 41, which identifies it as a "fairly difficult" text. Consequently, our system reduced the complexity of the evaluation text from "difficult" to "fairly difficult", while the human simplification was able to further reduce it to a "standard" level of difficulty.

3.3.3 Qualitative evaluation

Upon manual inspection of the differences between the human and the automatic simplification, we observed that the guidelines adhered to in the human simplification slightly differ from the ones we based the simplification rules of our system upon. For example, the human simplification contains commas and does not feature the *Mediopunkt*. Moreover, coordinate clauses with *und* ('and') are not split, explanations are not marked by indentations, and long simplified sentences are displayed on several lines.

[12]Recall that a lower LIX score points towards higher readability.

The biggest difference between the human and the automatic simplification is in lexical complexity. In the human simplification, many difficult words and expressions are replaced by simpler alternatives. For example, in the human simplification, the idiomatic expression *(jemanden) unter die Fittiche nehmen* ('to take (somebody) under the wings') is replaced with *sich kümmern* ('to take care of').

Our system segmented long words like *Meditations·techniken* ('meditation techniques') with the *Mediopunkt* and added Hurraki explanations for the words *Botschaft* ('embassy') and *Chef* ('boss'). Especially the explanation for *Botschaft* seems helpful; the human simplification does not explain this word. The human simplification introduces a new term *Schweizersportler* ('Swiss athletes'), which is not only long and possibly hard to read but also an incorrect compound word (correct: *Schweizer Sportler*).

For both simplifications, an exclamation mark was removed and sentences were split at dash signs and colons. An apposition was rephrased in a similar fashion in both texts. A final clause, sentential relative clause, and coordinative clauses were split in both simplifications, except for two *und* ('and') sentences in the human simplification, which were split only visually through line breaks. Since we had not implemented rules for rephrasing participle constructions, an occurrence of such a construction in the evaluation text remained unchanged by our system, while it was rephrased in the human simplification text.

In both the human and the automatic simplification, passive constructions and genitive attributes are resolved, although our system naturally returns more literal rephrasings. In one sentence, the dependency parser returned incorrect output, as a result of which a prepositional phrase could not be identified as agent. Apart from that, passive constructions were resolved correctly, even elliptic passive sentences. Genitive attributes were also rephrased correctly, with the exception of one case in which the wrong lemma was assigned by the parser, resulting in an incorrect dative form. In both simplifications, the past simple forms were changed to past perfect, with the exception of one sentence in the automatic simplification, where the predicate was not identified correctly by the parser.

While our system simply prints each output sentence on a new line, the human simplification text

splits long sentences at syntactic borders and displays them on several lines to improve readability.

3.4 Discussion

Our system still produces incorrect or unsimplified output for some sentences. For example, several subordinate clauses are not simplified because we have not found a general way of rephrasing them, e.g. conditional clauses, relative clauses, or clauses starting with *dass* ('that'). Furthermore, verbs with separable prefixes such as *ankommen* ('arrive') are not handled well, sometimes because the parser fails to identify the correct lemma, e.g., produces the lemma *kommen* ('come') instead of *ankommen* ('arrive'), and sometimes because our system does not incorporate a full-fledged grammar.

Moreover, our system provides limited support for reducing lexical complexity. However, even though it does not rephrase difficult vocabulary as readily as a human simplifier would, visual compound segmentation and the addition of explanations still aid in improving readability on the lexical level.

Overall, when applied to the evaluation text, our simplification system produced a readable simplified text. Especially on the syntactic level, the output of our system is comparable in complexity to the human simplification.

4 Conclusion and outlook

In this paper, we have reported on the development of a rule-based text simplification system for German. The rules underlying our system are based on linguistically motivated guidelines for transforming standard German into simplified German. Our system applies rules that perform simplification on the character, word, sentence, and text level and adjust the layout of the output.

We evaluated our system both quantitatively and qualitatively. With regard to quantitative assessment, our system was capable of reducing the complexity of the evaluation text from "difficult" to "fairly difficult" as measured by the LIX readability metric. The qualitative evaluation showed that our system does not reduce difficult vocabulary as readily as a human simplifier would; however, because compounds are segmented visually and explanations are added, readability is still increased at the lexical level. The syntactic complexity of the text output by our system was comparable to that of the human simplification.

In further developing our system, we intend to put more emphasis on lexical simplification, as even a text with short and simple sentences can be hard to read for inexperienced readers if it features high lexical density and contains difficult words. Apart from that, we are going to extend the syntactic simplification component in our system. As more resources for simplified German become available, we will be able to include synonym replacement, more elaborate explanation generation, and picture extraction in our system.

An automatic simplification system can only ever be as good as the rules it is built upon. In this sense, further theoretical research on simplified language is needed. In particular, the (potentially diverging) needs of the different target groups should be studied to greater detail. Preliminary efforts in linking the concept of simplified language to different levels of the Common European Framework of Reference for Languages (CEFR) are underway. For example, the *capito* network proposed three gradations of simplified language corresponding to the CEFR levels A1, A2, and B1.[13] Once these gradations become more formalized, it will be possible to implement them into automatic text simplification systems. As a result, these systems will be capable of producing different degrees of simplifications.

In further pursuing this work, collaboration with target readers will be most valuable for both designing rules and evaluating the system.

References

Sandra Maria Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, CA.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.

Bundesministerium der Justiz und für Verbraucherschutz. 2011. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung BITV 2.0). http://www.gesetze-im-internet.de/

[13] https://www.capito.eu/de/Angebote/Barrierefreie_Information/capito_Qualitaets-Standard/Guetesiegel_fuer_Leicht_Lesen/

bitv_2_0/BJNR184300011.html. Online. Last accessed March 3, 2016.

Jeanne Sternlicht Chall. 1958. *Readability: An appraisal of research and application*. Bureau of Educational Research, Ohio State University, Columbus, OH.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 1041–1044, Copenhagen, Denmark.

William Coster and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation (MTTG)*, pages 1–9, Portland, OR.

William Coster and David Kauchak. 2011b. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*, pages 665–669, Portland, OR.

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32:221–233.

Mariikka Haapalainen and Ari Majorin. 1995. GERTWOL und morphologische Disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference on Computational Linguistics*, Helsinki, Finland.

Katarina Heimann Mühlenbock. 2013. *I see what you mean: Assessing readability for specific target groups*. Ph.D. thesis, University of Gothenburg.

Inclusion Europe. 2009. Information für alle: Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht. http://www.inclusion-europe.org/images/stories/documents/Project_Pathways1/DE-Information_for_all.pdf. Online. Last accessed June 21, 2015.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA Annual Symposium Proceedings*, volume 2010, pages 366–370.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.

Christiane Maaß. 2015. *Leichte Sprache: Das Regelbuch*. LIT-Verlag, Berlin.

Kerstin Matausch and Annika Nietzio. 2012. Easy-to-Read and Plain Language: Defining Criteria and Refining Rules. http://www.w3.org/WAI/RD/2012/easy-to-read/paper11/. Online. Last accessed: November 13, 2015.

Netzwerk Leichte Sprache. 2009. Die Regeln für Leichte Sprache. http://www.leichtesprache.org/images/Regeln_Leichte_Sprache.pdf. Online. Last accessed June 21, 2015.

Gustavo Henrique Paetzold and Lucia Specia. 2015. LEXenstein: A Framework for Lexical Simplification. *ACL-IJCNLP 2015*, 1(1):85.

Evelina Rennes and Arne Jönsson. 2015. A tool for automatic simplification of Swedish texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 317–320.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology Conference*, pages 115–124.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*, pages 58–70.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *Computational Processing of the Portuguese Language. Proceedings of the 9th International Conference, PROPOR 2010*, pages 30–39, Porto Alegre, Brazil.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical Machine Translation with Readability Constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 375–386.

Martin Volk. 1999. Choosing the right lemma when analysing German nouns. In *Proceedings of the 11. Jahrestagung der GLDV*, pages 304–310, Frankfurt, Germany.

Willian Massami Watanabe, Arnaldo Candido Jr, Marcelo Adriano Amâncio, Matheus De Oliveira, Thiago Alexandre Salgueiro Pardo, Renata PM Fortes, and Sandra M Aluísio. 2010. Adapting Web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia*, 16(3):303–327.